Aplicación de Redes Neuronales al Filtrado de Documentos

A.C. Sergio A. Gómez[1], Lic. Laura Lanzarini[2]

Laboratorio de Investigación y Desarrollo en Informática[3]
Facultad de Informática - Universidad Nacional de La Plata

Resumen

El gran volumen de información disponible en formato electrónico supera ampliamente la capacidad de los usuarios, en lo que se refiere a procesamiento y acceso de información relevante.

La rama de los agentes inteligentes tiene como uno de sus objetivos brindar a los usuarios la ayuda necesaria para que puedan encontrar la información que buscan. Dichos agentes son capaces de "aprender" las preferencias de un usuario a través de la adquisición de perfiles de filtrado por medio de técnicas de *machine learning*.

Dado que cada documento puede ser caracterizado mediante un conjunto de patrones, es factible utilizar redes neuronales para determinar el perfil del usuario. Esta elección se basa en la capacidad discriminante de las redes neuronales y su amplia tolerancia al ruido en la información de entrada.

Esta línea de investigación busca estudiar y desarrollar arquitecturas de redes neuronales aplicadas al reconocimiento de patrones en documentos de texto.

Palabras Claves: Agentes inteligentes, Algoritmos de clustering, Filtrado de documentos, Redes neuronales.

1 Introducción

La gran cantidad de información disponible en la WWW aumenta continuamente. Esto ha llevado a la creación de diferentes motores de búsqueda que faciliten al usuario el acceso a la información deseada. Sin embargo, la cantidad de enlaces en los índices y la cantidad de referencias a documentos retornados por una consulta a Altavista o Lycos hacen inviable que el usuario tenga que analizarlos en forma individual, incrementando de esta manera su sensación de sobrecarga de información.

En el caso de las bases de datos, este problema está estandarizado (mediante consultas SQL), como también en el caso de grandes cantidades de datos numéricos (mediante el uso de técnicas de visualización gráfica). Por otra parte, en el caso del texto en lenguaje natural la situación es más compleja. La dificultad surge de la relación difusa entre la forma de los documentos (secuencias de caracteres) y su contenido semántico.

Además, el filtrado de documentos es complejo por las siguientes razones [Mostafa97]:

- *Dificultad de la representación*: Para operar eficientemente, los sistemas de filtrado de información (SFI) deben adquirir y mantener conocimiento preciso tanto de documentos como de usuarios. La naturaleza dinámica de los intereses del usuario hace muy difícil esta tarea.
- *Estocacidad del feedback*: El feedback de relevancia del usuario puede a veces parecer aleatorio desde el punto de vista del SFI. Esto puede ocurrir debido a varias razones. Primero, el usuario particular interactuando con el sistema puede tener necesidades inciertas o puede no ser muy discriminante de sus necesidades. También, en ciertas ocasiones, el feedback del usuario puede estar motivado por características particulares en los documentos que no son parte del esquema de representación subyacente. Luego, el feedback generado basado en esas características faltantes podría parecer aleatorio al SFI y éste sería incapaz de determinar qué causó tal feedback.

[1] Jefe de Trabajos Prácticos Dedicación Simple. Fac. de Informática. Universidad Nacional de La Plata. Becario de Entrenamiento Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. E-mail: sergiog@info.unlp.edu.ar.
[2] Profesor Adjunto Dedicación Exclusiva. Fac. de Informática. Universidad Nacional de La Plata. E-mail: laural@info.unlp.edu.ar
[3] Calle 50 y 115 1er. Piso, (1900) La Plata, Argentina, Tel./Fax +(54)(221)422-7707. http://lidi.info.unlp.edu.ar

- *Cambio de intereses del usuario*: Debido a razones profesionales o personales, los intereses de un usuario pueden evolucionar o reemplazarse por otros. Estos cambios pueden ocurrir a los largo de un período de corta duración o uno de larga duración.

El área de los agentes inteligentes se plantea como una forma viable de interactuar con los usuarios [Cheong96, Hunhs97, Pazzani97]. Los agentes se plantean como entidades autónomas y modulares que pueden aprender del usuario para satisfacer sus necesidades personales de información en forma automática.

Las redes neuronales aparecen como una solución al momento de plantear el modelo interno de los agentes de filtrado debido a su capacidad discriminante y su amplia tolerancia al ruido en la información de entrada [Freeman93].

2 Temas de Investigación y Desarrollo

La tarea de reconocimiento, clasificación y filtrado de documentos consta de las siguientes partes:

- *Obtención de los documentos a filtrar*: Los documentos a filtrar pueden obtenerse a partir de tres fuentes: i) a partir de un directorio de recursos (un documento HTML compilado por alguna persona con enlaces a documentos pertenecientes a un tema dado); ii) a partir de una URL y explorando el digrafo de la WWW con algún recorrido, y, iii) haciendo una consulta a uno o varios motores de búsqueda de la WWW.
- *Obtención de las características fundamentales de los documentos*: Consiste en obtener un conjunto de características o *features* de los documentos para obtener una representación de los mismos que sirva para usarlos de entrada a una red neuronal u otro mecanismo de clustering. En la literatura, generalmente la representación de los documentos está formada por un conjunto de términos con pesos asociados. Dichos términos pueden estar *stemmizados* previo filtrado de palabras stop.
- *Clustering de los documentos*: Los vectores de características de los documentos son agrupados mediante una red neuronal utilizando un criterio de similitud. Así, documentos similares relevantes (o irrelevantes) serán agrupados juntos. De esta manera, un perfil de usuario consistirá de un conjunto de clusters etiquetados como relevantes o irrelevantes. Un nuevo documento nunca antes visto será entonces asociado a un cluster existente (y de esta manera se predecirá su relevancia o irrelevancia) o disparará la creación de un nuevo cluster.

Los temas de investigación son los siguientes:

- Algoritmos de clasificación y clustering: Estudio de diferentes métodos de clustering no basados en redes neuronales y su aplicación en la clasificación de texto.
- Estudio de los diferentes modelos de redes neuronales que permitan resolver el problema de clustering de documentos.
- Lógica difusa. Inferencia neuronal difusa.

3 Resultados Obtenidos en los Trabajos Experimentales

La obtención de los documentos se hace a través de una interfaz basada en formularios HTML y guienes CGI que tienen capacidad de interactuar con los motores de búsqueda de la WWW. La obtención de los vectores de características a partir de los documentos HTML se hace realizando un "parsing" de los mismos teniendo en cuenta información de formato como encabezados y marcadores *meta*.

Los vectores de características se tomaron como entrada de diversos algoritmos de clustering como k-medias [Maravall94], red de contrapropagación, mapa autoorganizativo de Kohonen [Freeman93] y red de la Teoría de la Resonancia Adaptativa Difusa [Lavoie99]. Los algoritmos mostraron que son capaces de lograr una separación correcta de los documentos considerados en dos representaciones diferentes.

4 Líneas de Trabajo Futuro

Hay una variedad de direcciones en las cuales esta investigación puede orientarse en el futuro. Éstas se pueden separar en dos clases: representación y clasificación de documentos y estudio posterior e implementación de otros agentes y sistemas multiagentes.

Con respecto a la representación y clasificación de documentos, se puede decir:

- Se dejaron deliberadamente de lado otros formatos de documentos igualmente interesantes para el filtrado, como el formato de texto enriquecido (RTF), Postcript (PS), Abobe Portable Document Format (PDF), DVI, etc. Tampoco se consideraron los formatos comprimidos, por ejemplo: ARJ, ZIP, etc.
- Tampoco fueron tenidas en cuenta características tales de los documentos como el contenido multimedia, compuesto por fotos, películas y sonidos. Estas características bien podrían formar parte del perfil del usuario.
- Se puede estudiar la aplicación de algoritmos genéticos a arquitecturas evolutivas de redes neuronales [Michalewics92].

Con respecto a la implementación de agentes y sistemas multiagentes, se puede decir:

- Los experimentos realizados estuvieron orientados a un usuario simple. Una forma de extender este trabajo puede orientarse al filtrado cooperativo haciendo que varios usuarios den forma a un perfil de filtrado.
- Otra forma en la que se puede extender este trabajo es en la forma de la implementación de un sistema de filtrado multiagente en que, en vez de haber un único programa que realice el filtrado de los documentos, podría haber una multitud de los mismos realizando la tarea en mucho menor tiempo [Gómez99].

Bibliografía Básica

[Cheong96] Cheong, F. C. *Internet Agents: Spiders, Wanderers, Brokers, and Bots*. New Riders Publishing, 1996.

[Frakes92] Frakes, W.; Baeza-Yates, R. *Information Retrieval. Data Structures & Algorithms*. Prentice Hall, 1992.

[Freeman93] Freeman, J; Skapura, D. *Redes Neuronales. Algoritmos, aplicaciones y técnicas de programación*. Addison-Wesley/Díaz de Santos, 1993.

[Gómez99] Gómez S., *Una Taxonomía de Cambios para el Grafo de Espacios de Nombres Contextuales para Sistemas Multi-Agente en Contextos Múltiples.* V Congreso Argentino de Ciencias de la Computación. Tandil, 1999.

[Huhns97] Huhns, M.; Singh, M. *Readings in Agents*. Morgan Kaufmann Publishers, 1997.

[Lavoie99] Lavoie, Crespo y Savaria. *Generalization, Discrimination, and Multiple Categorization using Adaptive Resonance Theory*. IEEE Transaction on Neural Networks, Vol 10, nro. 4, Julio 1999.

[Maravall94] Maravall Gómez Allende, D. *Reconocimiento de Formas y Visión Artificial*. Addison-Wesley Iberoamericana, 1994.

[Michalewics92] Michalewics, Z. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer. Charlotte, 1992.

[Mostafa97] Mostafa, J.; Mukhopadhyay, S.; Lam, W.; Palakal, M. *A Multilevel Approach to Intelligent Information Filtering: Model, System, and Evaluation.* ACM Transactions on Information Systems, Vol. 15, No. 4, October 1997, Pp. 368–399. ACM 1046-8188/97/1000-0368.

[Pazzani97] Pazzani, M.; Billsus, D. *Learning and Revising User Profiles: The Iden-tification of Interesting Web Sites*. Machine Learning 27, 313-331. 1997. Kluwer Academic P.